DOCUMENT RESUME

ED 259 002 TM 850 371

AUTHOR Nandakumar, Ratna
TITLE An Application of Heckman's Correction for Specification Error.
PUB DATE Apr 84
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Admission Criteria; College Entrance Examinations; *Error of Measurement; Grade Prediction; Graduate Study; Higher Education; Law Students; Predictive Measurement; *Predictive Validity; *Regression (Statistics); Sampling; *Selective Admission; *Statistical Bias
IDENTIFIERS *Heckman (J J); Law School Admission Test; Specification Bias

ABSTRACT

Heckman's correction for regression in selected samples for predictive validity studies was applied to a large data file on 7,984 law school applicants. Data included ethnic group, sex, socioeconomic status, undergraduate degree, school, scores on the Law School Admission Test (LSAT), writing ability, undergraduate grade point average, and age. The final selection criteria were not known. Data on the 1,845 applicants who were accepted included year of entrance, sex, date of birth, undergraduate grade point average, grade point average for first year of law school, undergraduate college, LSAT score, writing ability, ethnic group, and age. The grade point average for the first year of law school was estimated and compared, with and without using Heckman's correction factor. Three different methods were used to choose the variables for student selection and grade prediction: (1) the same set of variables were used for both selection and prediction; (2) a subset of variables used for selection were used for prediction; and (3) two completely different sets were used. It was found that the relation between the variables used for selection and prediction could affect the accuracy of prediction. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED259002

An Application of Heckman's Correction for Specification Error

Ratna Nandakumar
University of Illinois
150,CRC
Champaign, Il 61820

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Nandakumar, Ratna

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1984

TM 850 371

ABSTRACT

In this papaer an attempt is made to apply Heckman's procedure for correction of specification error in selected samples. A large law school data set has been used in the study. The first year average in law school is estimated and compared with and without using Heckman's correction factor. Three different ways are considered to select the variables for selection and prediction: 1. The same set of variables are used both for selection and prediction. 2. A subset of variables used for selection is used for prediction. 3. Completely different sets of variables are used for selection and prediction. It was found that the relation between the variables used for selection and prediction can affect the accuracy of prediction.

Introduction

Predictive validity studies take place in situations where individuals are selected on some basis to meet a specified criterion. If a test score is used as a basis of selection, predictive validity is obtained by comparing test scores with the criterion variable considered to provide a measure of the characteristic in question. For example, in colleges and universities students are selected on the basis of test scores which predict their academic success or a firm hires people on the basis of several factors that predict the job success. Predictive validity studies are used to see if tests used for selection predict the performance or a specified criterion. In the example of college students one would want to know if students selected on the basis of test scores perform better than unscreened group of students i.e., to see if test used for selection predicts the performance. Decision makers use predictive validity results for placement of people to different treatments according to the test scores.

In such studies we often come across situations where the selection of units into the sample is not random. In such situations it is important to model the selection process (the process by which the observed units are selected into the sample). This is typical when the students are admitted into a college. The Restrictive samples may also occur due to

attrition when the individuals voluntarily participate in the program (self selection) and in longitudinal studies. Failure to recognize the sample selection and analysing the selected samples as though it were random can have serious consequences such as biased and inconsistent estimates of the parameters. As said in Linn(1982), "Fortunately randomization is not the only approach to obtaining unbiased estimates of regression." What is important is to avoid bias by gaining full knowledge about the selection process and model it. Selectivity problems have been discussed in various contexts by many authors, e.g. Gonan (1974), Lews (1974), Heckman (1974, 1977), Goldberger (1981) and Linn (1982).

Linn(1982) showed how the selection process can produce overprediction results for minority groups in two different cases of populations. Case(1) There is no overprediction in the population prior to selection;i.e., the regression equations of majority and minority groups are equal. Case(2) The majority group regression equation underpredicts the average minority group performance in the population prior to selection. For each of these two cases, three possible cases of minority group selection is considered: (a) random selection or no selection. (b) selection on the basis of third variable U defined as for the majority group selection. (c) selection on the basis of U',which places less weight on X for minority-group than for majority-group members.

It was found that overprediction was the consequence of the selection process for all three minority group selection situations for Case(1). For Case (2) there is overprediction in the selected sample for Cases (a) and (b) but underprediction for Case(c).

It was also found that amount of overpredition will be larger for highly selective situations. As the selection is higher, the standard deviation of the predictors is smaller and the degree of overprediction is high.

In this paper Heckman's correction for regression in selected samples is applied to a large law school data file. The applicant file consists of data for 7984 subjects on variables:ID, decision made(admit/reject),ethnic group, sex, socioeconomic status, undergraduate degree, school, scores on LSAT, and Writing ability of last three attempts(if any), UGPA and age. From 7984 subjects school accepted 1845 subjects.No information is available on criterion for selection. The accepted file consists of data for 1845 subjects on the following variables: ID, year of entrance, sex, date of birth, UGPA, first year average in law school(FYA), undergraduate college code, scores on LSAT and Writing ability of last three attempts(if any), ethnic group and age.

The applicant file was used for Probit analysis to obtain the parameters of equation (8) (explained below) and the results are used to obtain Heckman's correction factor. The accepted file was used for obtaining the least-squares estimates of regression equations : (1) regression equation with Heckman's correction factor. (2) regression equation without Heckman's correction factor.

Method and application

Consider the simple regression of Y on X in the unselected population,

(1) $$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ is uncorrelated with X.

If one has random samples of observations on Y and X, one can obtain unbiased estimates of β_0 and β_1 by ordinary least squares (OLS). If the sampling is non random, the units are selected on the basis of a possibly unobservable variable U. The units are selected into the sample if U exceeds a threshold value (say $U > 0$). In this case the residual is no longer independent of X, and so the estimates will be biased as a function of the sampling process. As stated by Berk, Ray & Cooley (1982), there are problems of external validity and internal validity. First, the regression line for the original population does not correspond to the regression line for the

selected population. The regression parameters differ depending upon the data available. This is a problem of external validity. Second, the error term is correlated with the regressor and this is a problem of internal validity. The estimates of regression coefficients are biased and the linear regression model is the wrong model even in the selected population. This is conceived by Heckman as specification error in the original model which has no parameter representing the selection process. Conventional formulas for correcting the restriction of range may not be appropriate due to the limitation of underlying assumptions, i.e. linearity and homoscedasticity. Furthermore,application of traditional range restriction formulas requires that all variables,which contribute to the selection process be included in the analysis,but this is rarely possible since the percise basis of selection is often unknown.

When U = X, i.e., when the selection is based explictly on the independent variable, there is zero probability of selecting population units to the left of this value. In this case if the units to the right of U are selected at random, then there is no specification error and OLS gives unbiased estimators of β_0 and β_1 . However correlation estimates between X and Y are affected because of reduced variance of X in the selected population.

When U = Y, i.e., when the selection is based explicitly on dependent variable, there is no probability of selecting population units below (above) this value into the selected population. In this case the error term will be correlated with X in the sample. The mean of error values will be higher for units with smaller X values. If U = 0, this corrsponds to the Tobin(1958) model (correlation of limited dependent variables). In this case OLS gives biased estimates of the slope β_1, which is biased downwords and is inconsistent for large samples. The regression line in the selected population will have a smaller slope and higher intercept. Under the assumption of multivariate normality the relations due to Goldberger (1981), between the regression paramaters in the original and selected populations can be written as :

(2) $\beta^* = c\beta$,

(3) $\alpha^* = (1 - c\rho^2)\mu_y^*$,

(4) $\rho_i^2 = c\rho^2$, where

(5) $c = \theta / [1 - \rho^2(1-\theta)]$ and

(6) $\theta = V^*(y) / \sigma^2$

where asterisks indicate the parameters in the selected population.

Equations (2) and (4) show that the slope and the multiple correlation coefficients in the selected population are proportional to the slope and correlation coefficients in the original population, the constant of proportionality being the constant c . The impact of selection is therefore represented by constant c , which in turn depends on θ . From Equation (6), θ is the ratio of variance of the dependent variable Y in the selected population to the variance in the original population.

From Equations (5) & (6), assuming $0 < \rho < 1$, it can be seen that, when $\theta > 1.0$, the constant $c > 1.0$ and the regression coefficient β^* in the selected population is inflated. When $\theta < 1.0$, the constant $c < 1.0$ and the regression coefficent in the selected population is deflated. When $\theta = 1.0$, the regression coefficients are equal in both the populations. Therefore the crucial point is the relation between the variance of Y in the selected population and the original population.

When $\rho = 1$, $c = 1$ and there is no effect due to selection. For $\rho = 0$, $c = \theta$. But in reality ρ will not take these extreme values. The intercept in the selected population is also distorted.

The sample drawn from the selected population, therefore produces inconsistent estimates of the regression parameters in the original population, with a degree of inconsistency depending on the value of the constant c . For example, when c = 0.5 the estimated regression coefficinets will be approximately half of the original values, if the usual linear regression model is applied and hence the external and internal validities are in doubt.

When U is a third variable used as a basis for selection OLS in the regression of Y on X will be different for different subpopulations and there will be a decrease in the slope and a concommittant increase in the intercept. For all cases Dunbar (1982) has illustrated the results by simulation methods for normally distributed variables.

In the case of U used as a (unobservable) third variable as the basis of selection, Heckman treats the regression as a two stage equation model. One equation describes the relationship between Y and X and the other describes the selection process. Equation (1) can be replaced by two equations:

(7) $$Y = \beta_0 + \beta_1 X + \epsilon ,$$

(8) $$U = G_0 + G_1 X + \delta$$

We assume that in the total population the joint distribution of ϵ and δ is normal and independent of X with means zero and

the covariance matrix of ϵ and δ is given by

$$\underset{\sim}{\Sigma} = \begin{pmatrix} \sigma_{\epsilon\epsilon} & \sigma_{\epsilon\delta} \\ \sigma_{\epsilon\delta} & \sigma_{\xi\xi} \end{pmatrix}$$

where $\sigma_{\epsilon\delta}$ = covariance between ϵ and δ, $\sigma_{\epsilon\epsilon}$ = variance of ϵ and $\sigma_{\xi\xi}$ = variance of δ. Consider the regression of ϵ on δ,

(9) $$\epsilon = \omega\delta + \eta$$

η is uncorrelated with δ and $\omega = \sigma_{\delta\epsilon} / \sigma_{\xi\delta}$.

W' hout loss of generality it is assumed that $\sigma_{\xi\xi} = 1$ so that

T.e regression of Y on X for the selected population units with $U > 0$ is

(10) $$E(Y|X, U>0) = \beta_0 + \beta_1 X + E(\epsilon | U>0).$$

But from using equations (8) and (9) in (10) we obtain

(11) $$E(Y|X, U>0) = \beta_0 + \beta_1 X + \omega E(\delta | \delta > -G_0 - G_1 X)$$

Let

(12) $\lambda(X) = G_0 + G_1 X$,

and $f(\lambda)$ is defined by

$$f(\lambda) = E(\delta | \delta > -\lambda(x)).$$

Equation (11) can be written as

(13) $$E(Y|X, U>0) = \beta_0 + \beta_1 X + \omega f(\lambda)$$

It is clear that OLS regression of Y on X will not give a consistent estimate of β_1 unless $\omega = 0$ i.e, $\sigma_{\delta\epsilon} = 0$

Let P(z) denote the probability distribution of δ and p(z) denote the density function of ε. We assume p(z) to be symmetric about zero, so that p(-z) = p(z). Then,

$$\text{Prob}(\delta > -\lambda) = P(\lambda), \quad \text{and}$$

$$(14) \qquad f(\lambda) = E(\varepsilon \mid \delta > -\lambda) = \frac{1}{P(\lambda)} \int_{-\infty}^{\lambda} z\, p(z)\, dz,$$

when δ has the standard normal distribution, let $\phi(z)$ and $\Phi(z)$ denote the density and distribution function of δ. (14) can be written as:

$$f(\lambda) = \frac{\phi(\lambda)}{\Phi(\lambda)}.$$

In Equation (13), if the covariance between the error terms in the Equations (7) & (8) is zero, the regression weight associated with $f(\lambda)$, ω is 0 and the effect of incidental selection disappears. The function $f(\lambda)$, which is a monotonically decreasing function of λ, represents the probability that an observation is selected into the sample. If $f(\lambda)$ is large, the likelihood of inclusion into the sample is large and vice versa. Also, $f(\lambda)$ is the expectation of the error term in the selection equation after selection. After selection, $f(\lambda)$ in the selection equation is nonzero and if ω is not equal to zero, contributes for incidental selection in Equation (13) and is correlated with the regressor.

In the first step of Heckman's procedure $\underset{\sim}{\gamma}$ is estimated on the full sample by maximum likelihood probit analysis with the dependent variable Y coded '1' if an individual is selected and '0' if an individual is not selected. The ~~predicted~~ estimated values from Equation (8) are then used to construct $f(\lambda)$. In the second step OLS is applied to Equation (13), with the estimated $f(\lambda)$ as an additional predictor. The resulting regression coefficients and the intercept from Equation (13) are consistent and unbiased. If the regressors in both the Equation (7) and (8) are very similar, it is common to find high multicollinearity between $f(\lambda)$ and other regressors in Equation (13) and this makes it difficult to determine the importance of selection effects.

The conditional mean and variance of Y are given by:

$$(15) \quad E(Y|X, U>0) = \beta_0 + \beta_1 X + \omega f(\lambda),$$

$$(16) \quad Var(Y|X, U>0) = Var(\epsilon|U>0)$$
$$= \sigma_{\epsilon\epsilon} - \omega^2 f(\lambda)[\lambda + f(\lambda)].$$

Least-squares solution for parameters of Equation (13) using the data from only the selected sample gives unbiased and consistent estimators of the parameters provided the probit model is correctly specified. The standard errors for estimated values in Equation (13) are larger than those that would be obtained if the model were applied to the entire

sample. Heckman notes that the conventional formulas for standard error applies only when $\rho(\epsilon,\delta)=0$. Otherwise the conventional standard errors are underestimates of true standard errors. From Equation (14) it can be seen that $f(\lambda)$ is a nonlinear function of X and hence the true regression of Y on X is nonlinear. Also from Equation (16), since ω is not equal to zero and $f(\lambda)$ is not a constant, the conditional variance of Y is heterogenous. As stated above inclusion of $f(\lambda)$ as an additional predictor in the true regression of Y on X introduces considerable amount of collinearity. This added collinearity also contributes to the instability of the OLS estimators.

Generalizations to the multivariate case can be made in a straightforward manner to p arbitrary predictors;
Let B' and G' be the vectors of regression coefficients. The model can be written as:

$$Y = \underset{\sim}{B}'X + E, \text{ and}$$

$$U = \underset{\sim}{G}'X + D,$$

where

$$Y = \begin{cases} \text{observed if } U > 0 \\ \text{not observed otherwise} \end{cases}$$

Main analysis and results

The law school applicant and accepted files were used to illustrate Heckman's procedure for estimating the parameters of regression equations for a selected sample. The applicant file consists of 7984 cases and the accepted file consists of 1845 cases. Three main types of analysis were performed. The variables used for selection in probit analysis and the variables used for predicting first year average in law school are differently selected in each of the three cases. In Case (1) the variables (explained below) UGPA and ALSAT were used as basis for selecting students for admission into law school and the same variables were used along with Heckman's correction factor for predicting first year average in law school. In Case (2), the first principal component was used as basis for selection, which was obtained with variables (explained below) ALSAT, UGPA, AWA, SEX and RACE. The variables UGPA and ALSAT were used along with Heckman's correction factor for predicting the first year average. In Case (3) variables AWA, SEX and RACE were used for selection and a completely different set of variables i.e., UGPA and ALSAT were used for predicting the first year average. It is found that the relation between the sets of variables used for selection and prediction can influence the accuracy of prediction. Each case is briefly described and the results are summarised below.

In Case (1) undergraduate GPA (UGPA) and average LSAT (ALSAT) were used as variables as basis of selection in multiple probit analysis (using the statistical package program SOUPAC) with applicant file. The resulting estimates are used to get $f(\lambda)$ by a subroutine called MSMRAT from international mathematical and statistical library (IMSL). First year average in law school is predicted with and without the correction factor $f(\lambda)$: 1) UGPA, ALSAT and $f(\lambda)$ as the predictors and 2) UGPA and ALSAT as predictors for the accepted file. This regression was performed with the statistical package SPSS. The results are summarized in Tables 1 and 2.

The top part of Table 1 shows the selection equations applied to the full applicant population file of 7984 individuals with dummy variable Y taking the value 1 if selected and 0 otherwise. It appears that individuals with higher UGPA are more likely to be selected into the law school. Table 2 gives the regression estimates with and without lambda i.e., B(Heck) & B(OLS) respectively for the accepted file. Lambda represents the probability of an individual being slected into the law school. In this case the correction for the sample selection bias does not make much of a difference in spite of the strong selectivity, where only one quarter of applicant population has been selected. The differences in R 's in two cases is negligible. As can be expected because of high selectivity problem, ordinary regression results are

biased. The corrected regression weights are higher than the uncorrected ones. For predicting first year average in law school, variable ~~ALSAT~~ UGPA appears to be a better predictor than others. Even though the Lambda value is not statistically significant, the direction of the impact seems reasonable. Individuals who have high probability of being selected are likely to have high first year averages. In this example, the regressors used for selection and prediction equations are exactly the same. This adds to the problem of multicollinearity that follows even if the regressors are different in two equations and hence contributes to the reduction of selection effect. On the other hand, this also implies that the correlations between the error terms in the two equation model, i.e., $\omega = .3394$ is too small for significant selection effect, but indicates some selection effect. Berk, Ray & Cooley (1982) note that the cross correlation between the error terms often turn out to be small under properly specified models. Consequently the selection artifacts will also be small.

Summary of results for Case (1)

TABLE 1

1. Selection equation (1=applicant selected)

Variable	Max lik est	Std.wt
UGPA	1.1282	1.61E-02
ALSAT	0.0052	0.02E-07
Intercept	-7.9450	

N = 7984
No. selected = 1845
Std.wt = Standardised weight

TABLE 2

Results in the selected sample

	Corrected		Uncorrected	
Variable	B(Heck)	p	B(OLS)	p
UGPA	1.1342	.354	0.8214	0
ALSAT	0.0130	.019	0.1161	0
LAMBDA	0.3394	.797		
Const	-5.0654	.595	-2.6217	.43

(1) Regression of FYA on UGPA, ALSAT, LAMBDA :

$R = 0.5121 \qquad R^2 = 0.2622$

$Y_h = -5.0656 + (1.1342)\ UGPA + (0.0130)\ ALSAT + (0.3395)\ LAMBDA$

(2) Regression of FYA with UGPA, ALSAT :

$R = 0.5121 \qquad R^2 = 0.2622$

$Y_{ls} = -2.6217 + (0.8215)\ UGPA + (0.0116)\ ALSAT$

$r(Y_h, Y_{ls}) = 0.9999$

In Case (2) the first principal component was computed with the variables which included the average writing ability (AWAB), ALSAT, UGPA, sex, and RACE from the original accepted file. A pseudo accepted file was created from the original accepted file with the top 500 cases on the first principal component. The original accepted file with the additional decision vector (1 if selected, 0 otherwise) was used for probit analysis and the results were used to get $f(\lambda)$, as in the Case (1) using SOUPAC and IMSL packages. The regression is performed on the pseudo accepted file to predict first year average with predictors 1) UGPA, ALSAT, $f(\lambda)$ and 2) UGPA, ALSAT.The results are summarized in Tables 3 and 4.

From Table 3,it can be seen that individuals with higher scores of ALSAT are more likely to be selected, although UGPA,AWA,and RACE also contribute very highly for selection. Table 4 gives the regression estimate with and without the correction factor which were computed for individuals in the pseudo selected file. For this case, UGPA contributes significantly for predicting the first year average. As can be expected as a result of double selection, the variability of predictors is very low and so R values are very small.The correction factor for sample bias is nearly zero.It does not make any difference for prediction whether LAMBDA is used or not.

Summary of results for Case (2)

TABLE 3

Principal component:

PC = (0.8870)ALSAT + (.8062)AWA + (0.7986)RACE + (0.4433)UGPA
- (0.0699)SEX

1. Selection equation (1=applicant selected)

Variable	Max lik est	Std.wt
ALSAT	42.98	192.65E+15
UGPA	32.51	147.29E+15
AWA	37.87	134.79E+15
SEX	- 1.50	2.14E+15
RACE	33.07	144.58E+15
Constant	-32577	

n = 2000
NO. selected = 500
Std.Wt = Standardised weight

TABLE 4

Results in the selected sample

	Corrected		Uncorrected	
Variable	B(Heck)	p	B(OLS)	p
UGPA	0.3834	.016	0.3834	.016
ALSAT	0.0016	.919	0.0079	.674
LAMBDA	0.00014	.690		
Constant	5.831	.602	1.4753	.542

(1) Regression of FYA on UGPA, ALSAT, LAMBDA :
R = 0.1491 R^2 = 0.02224
Y_h = 5.8310 + (0.3834) UGPA + (0.0016) ALSAT +
(0.0001)LAMBDA

(2) Regression of FYA with UGPA, ALSAT :
R = 0.1481 R^2 = 0.02192
Y_{LS} = 1.4753 + (0.3834) UGPA + (0.0079) ALSAT
$r(Y_h, Y_{LS})$ = 0.8324

In Case (3) AWAB, SEX and RACE were used as variables which formed the basis of selection in probit analysis using the original applicant file. The results were used to get $f(\lambda)$ as in the previous cases. Regression is performed to predict first year average in law school with predictors 1) UGPA, ALSAT, $f(\lambda)$ 2) UGPA, ALSAT for the original accepted file. The results are summarized in Tables 5 and 6.

The top part of Table 5 shows that RACE contributes for the selection much more highly than AWA and SEX. Whites are coded as '1' and non-Whites are coded as '0'. The results shows that non-Whites are more likely to be selected than Whites. Males are coded as '1' and females are coded as '2'. Males are more likely to be selected than females. In this instance the correction for sample bias is highly significant and contributes for prediction. UGPA contributes for prediction much more highly than other variables.But surprisingly R values are nearly the same in corrected and uncorrected cases. One possible explanation could be the values of the intercept. In the overall prediction, higher values of the intercept in the uncorrected case may be compensating for the selection effect. High values of LAMBDA also implies that the cross correlation of error terms in the two equation are highly correlated, introducing specification error.

BEST COPY AVAILABLE

Summary of results for Case(3)

TABLE 5

1. Selection equation (1=applicant selected)

Variable	Max lik est	Std.wt
AWA	0.0476	0.1163E-05
SEX	-0.0617	0.0396E+02
RACE	-0.6146	0.7746E-02
Constant	-0.2974	

N = 7984
No. selected = 1845
Std wt = Standardised weight

TABLE 6

Results in the selected sample

	Corrected		Uncorrected	
Variable	B (Heck)	p	B (OLS)	p
UGPA	0.7659	.000	0.8215	.000
ALSAT	0.0125	.000	0.0116	.000
LAMBDA	0.7379	.000		
Constant	-3.9515	.000	-2.6216	.000

(1) Regression of FYA on UGPA, ALSAT, LAMBDA :

$R = 0.52001$ $R^2 = 0.27041$

$Y_h = -3.9515 + (0.7659)\ UGPA + (0.0125)\ ALSAT + (0.7379)\ LAMBDA$

(2) Regression of FYA with UGPA, ALSAT :

$R = 0.51205$ $R^2 = 0.26220$

$Y_{LS} = -2.6217 + (0.8215)\ UGPA + (0.0116)\ ALSAT$

$r(Y_h, Y_{LS}) = 0.9883$

Overall discussion of all three cases:

In all three cases the difference in multiple R 's, with Heckman's correction factor and with ordinary regression were negligible. In case(2), because of selection on top of selection, the variability of predictors was reduced very much and so the R value is very small. In the other two cases, although R values are quite significant, because of high correlation between variables used as basis of selection and also used as predictors in the subsequent regression, the regression weights are very small.

For all three of the above cases scatterplot of predicted scores using correction factor "H predicted FYA" versus predicted scores without the correction factor "LS predicted FYA" are plotted in Figures (1), (2) and (3) for about 100 randomly chosen cases. As can be seen the two predicted values are highly correlated in all the three cases, taking values 0.99, 0.83 and 0.99 respectively. So in this particular situation, it would rarely make any difference which equation is used for prediction purposes. However,it may make difference which equation is used for prediction for particular individuals. For example, if we consider the top 60 people for selection, and set cut off lines as shown in Fig(3), it can make a difference for the individuals in regions D and for one individual in region B, which equation is used for selection.

Ordinary least-squares equation rejects individuals in region B and accepts individuals in region D, whereas Heckman's corrected regression equation rejects individuals in region D and accepts individuals in region B.

Conclusion

The selection artifacts are pervasive in applied research. Any data can be viewed as a selected subset from some larger population. The solution of selection problems is based upon the proper modelling of the selection process. In situations of explicit selection, under multivariate normality, results of Goldberger (1981) can be used for correcting the selection artifacts. Goldberger's results are robust when the assumption of multivariate normality is not reasonable. In problems of incidental selection there are no straightforward corrections even under multivariate normality assumptions. The impact of selection depends much upon the correlation between the error terms in the two equation model. Berk, Ray and Cooley (1982) note that correlations of near zero mean that incidental selection effects are minor and correlations over .80 are grounds for serious concern. Also there is the problem of multicollinearity that can influence this correlation. It is assumed that the errors in the two equations ϵ and δ have a bivariate normal distribution. Alternatively one can assume a) rectangular distribution, b) logistic distribution, c) errors are linearly related. Again Berk, Ray and Cooley (1982) conclude that it makes little difference in practice which of the estimators one uses.

In the foregoing discussion, the selectivity problem in one group was considered. The method can be extended when there are more than one non-equivalent groups. Muthen (1981) has shown the statistical and computational ways of analyzing selective samples by modeling the selection process in each group. A simulation study by Muthen (1981) has been shown to illustrate the failure of ANCOVA to show the significance of the treatment effect in two groups due to selectivity problem.

In order to use Heckman's procedure for estimating the unbiased estimates in selected samples the data requirement is that the information on X used in Equation (3) be known for the entire unrestricted sample. If the data is not available on unselected applicant sample the method of estimating the unbiased parameters is given in Craig (1983). Research on the violation of the assumption of bivariate normal distribution of ϵ and δ is limited. Goldberger (1980) notes that when ϵ and δ are not bivariate normal, the results are biased but less biased than when OLS is used ignoring the selection process.

REFERENCES

Berk, R.A., Ray, S.C. & Cooley, T.F. Selection biases in sociological data. Project Report, National Institute of Justice (Grant No. 80-IJ-CX-0037). University of California, Santa Barbara, 1982.

Barnow, Burt s., Cain, Glen c., & Goldberger, Arthur S. Issues in the analysis of selectivity bias.In Ernst W. Stromsdorfer and George Farkas (Eds.), Evaluation Studies Review Annual , Vol. 5. Sage Publications: Beverly Hills/ London, 1980.

Bengt, M. and Joreskog, K. G. Selectivity problems in quasi experimental studies. Paper presented for the conference on Experimental Research in Social Sciences ,Gainsville, Florida, January, 1981.

Dunbar, S. Corrections for sample selection bias. Ph.D Thesis, 1982

Goldberger, A.S. Linear regression after selection. Journal of Econometrics ,1981, 15 , 357-366.

Gronau, R. Wage Comparisons - A selectivity bias. Journal of Political Economy ,82 , 1119-1144, 1974.

Craig, A. Olson and Brian, E. Becker. A proposed technique for the treatment of restriction of range in selelction validation. Psychological Bulletin ,1983, 93 , 137-148

Heckman, J. J. Sample selection bias as a specification error. Econometrica ,1979, 47 ,153-161

Lewis, H. G. Comments on selectivity biases in wage comparisions. Journal of Political Economy ,1145-1155,1974

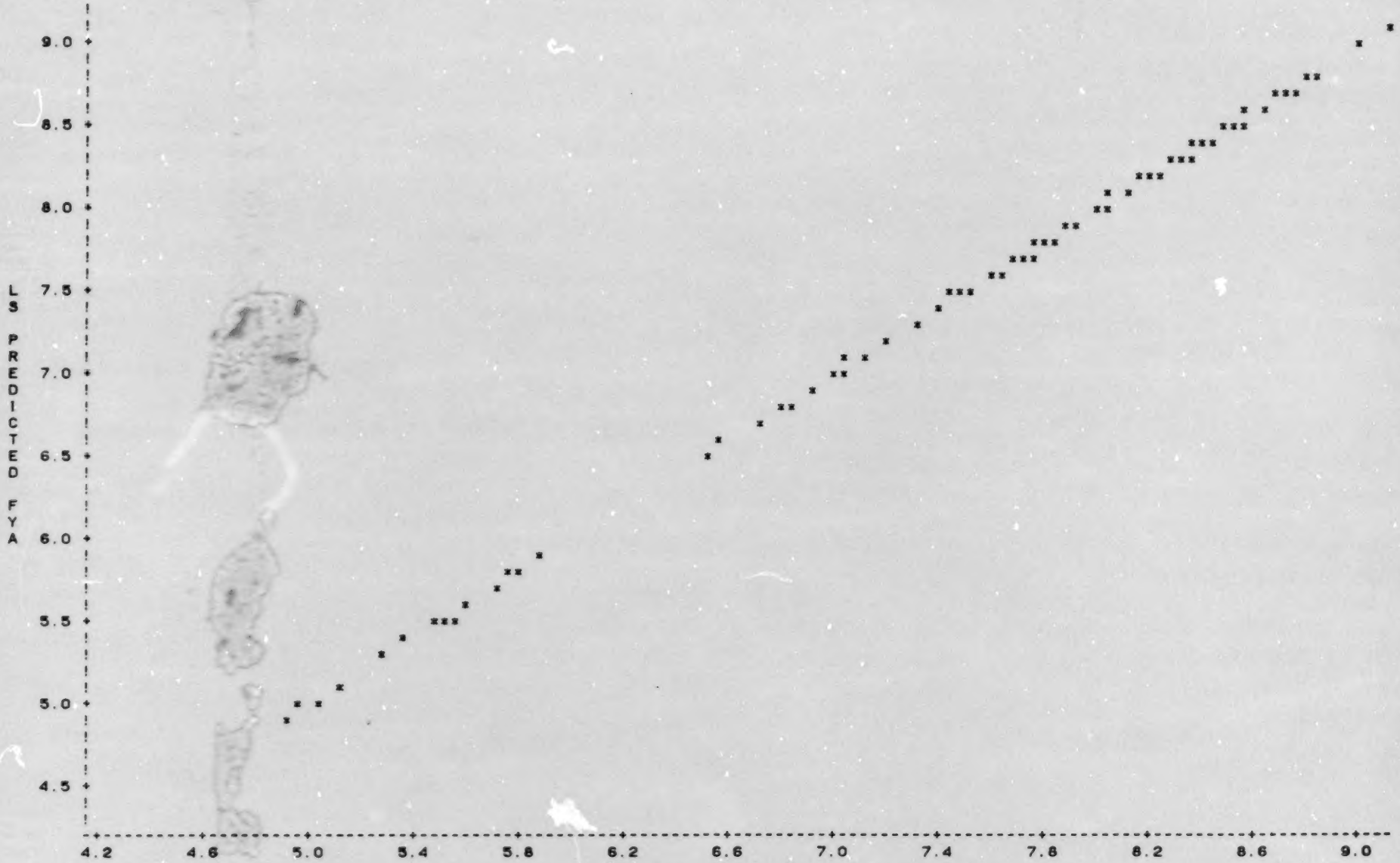
Linn, R. L. Predictive bias as an artifact of selection procedure. Paper prepared for "Advances Psychometric Theory: A Festschrift for Frederic M. Lord," Princeton, NJ: Educational Testing Service, May, 1982.

Tobin, J. Estimation of relationships for limited dependent variables. econometrica , 1958, 26 , 24-36

Thorndike,R.L. Educational Measurement. American Council on Education, Washington,D.C.,1971.

PLOTS FOR CASE 1

PLOT OF YOLS1*YHECK1 SYMBOL USED IS *



NOTE: 1 OBS HAD MISSING VALUES 31 OBS HIDDEN

Figure 1

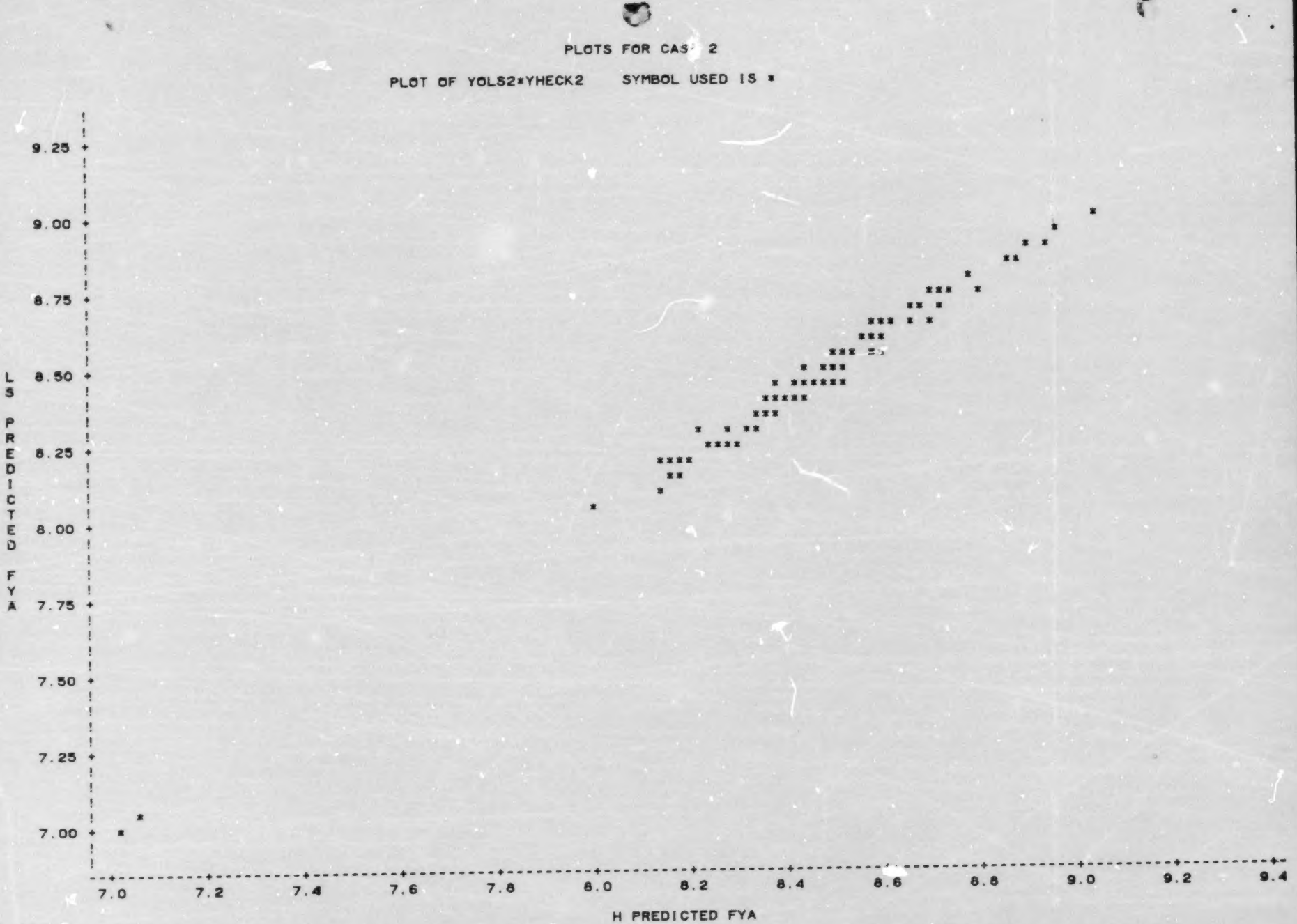

PLOTS FOR CASE 2
PLOT OF YOLS2*YHECK2 SYMBOL USED IS *
LS PREDICTED FYA
9.25
9.00
8.75
8.50
8.25
8.00
7.75
7.50
7.25
7.00
7.0
7.2
7.4
7.6
7.8
8.0
8.2
8.4
8.6
8.8
9.0
9.2
9.4
H PREDICTED FYA
NOTE: 36 OBS HIDDEN

Figure 2

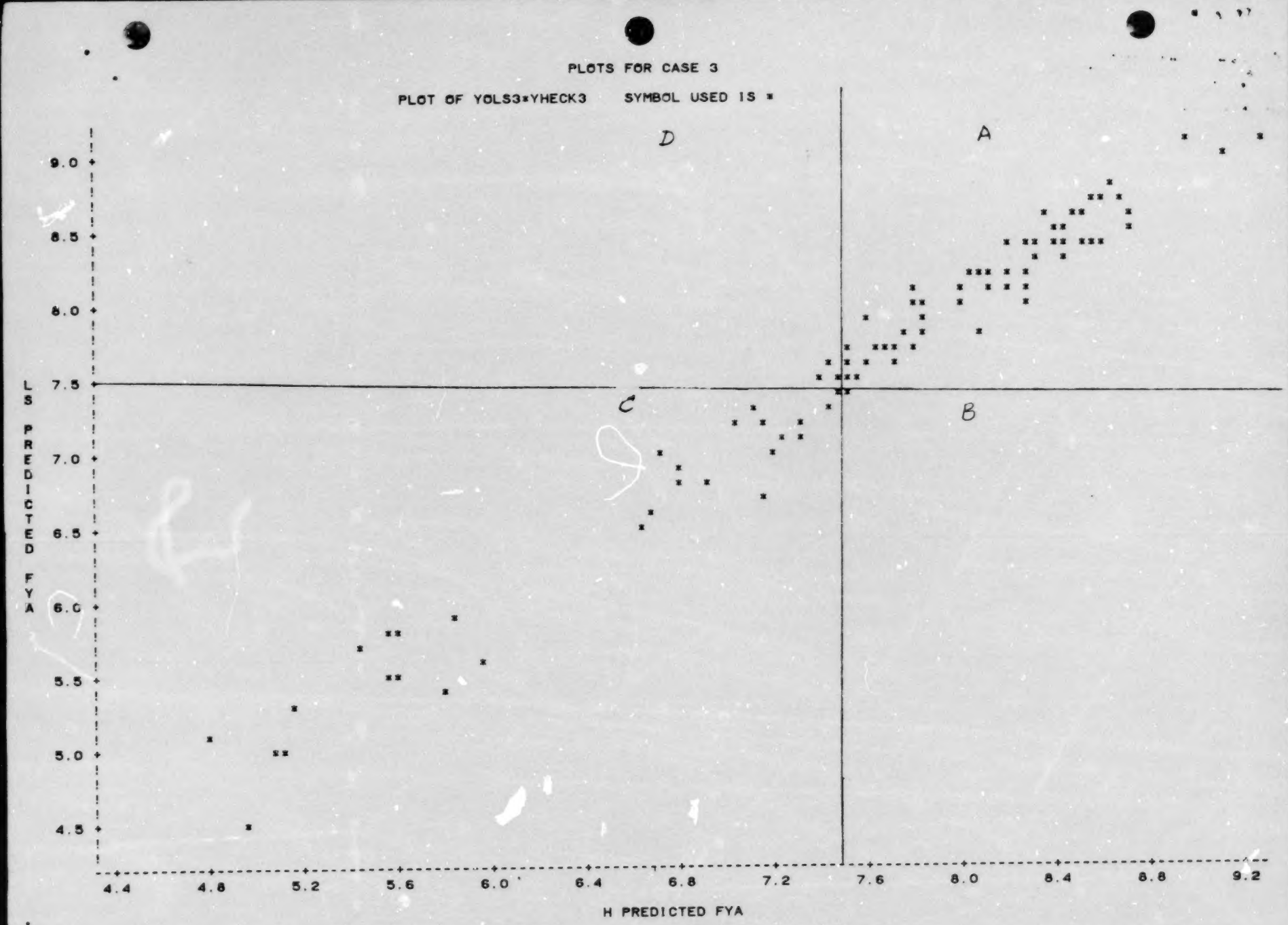

PLOTS FOR CASE 3
PLOT OF YOLS3*YHECK3 SYMBOL USED IS *
D
A
C
B
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
LS PREDICTED FYA
4.4
4.8
5.2
5.6
6.0
6.4
6.8
7.2
7.6
8.0
8.4
8.8
9.2
H PREDICTED FYA
NOTE: 10 OBS HIDDEN

Figure 3